Exploring IT Cost Components – How to Maximize your IT Investments

Ray Jones

Vice President, System z Worldwide Software Sales,
IBM Software Group

Many Cost Components

80:20 rule helps to achieve reasonable results in a short time

Components

Hardware

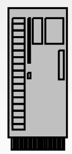
Software

People

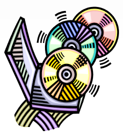
Network

Storage

Facilities



List vs Discounted
Fully configured vs. basic, Prod. vs. DR
Refresh / upgrade, Solution Edition…



IBM and ISV, OTC and Annual maint (S&S)
MLC, PVU, RVU, ELA, core, system

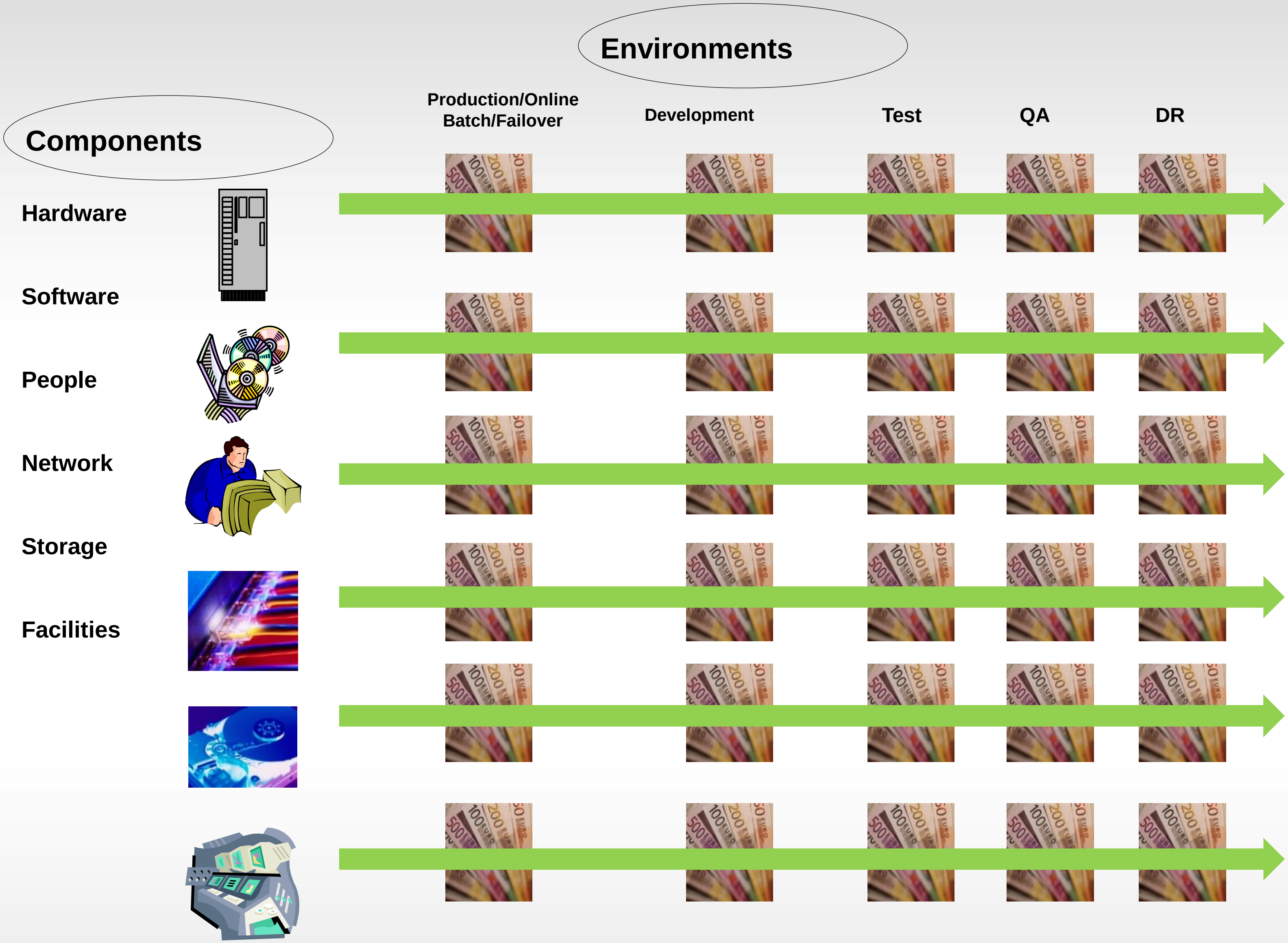
FTE rate, in house vs. contract

Adapters, switches, routers, hubs
Charges, Allocated or apportioned, understood or clueless

ECKD, FBA, SAN, Compressed, Primary, secondary
Disk (multiple vendors), tape, Virtual, SSD

Space, electricity, air cooling, infrastructure including UPS and generators, alternate site(s), bandwidth

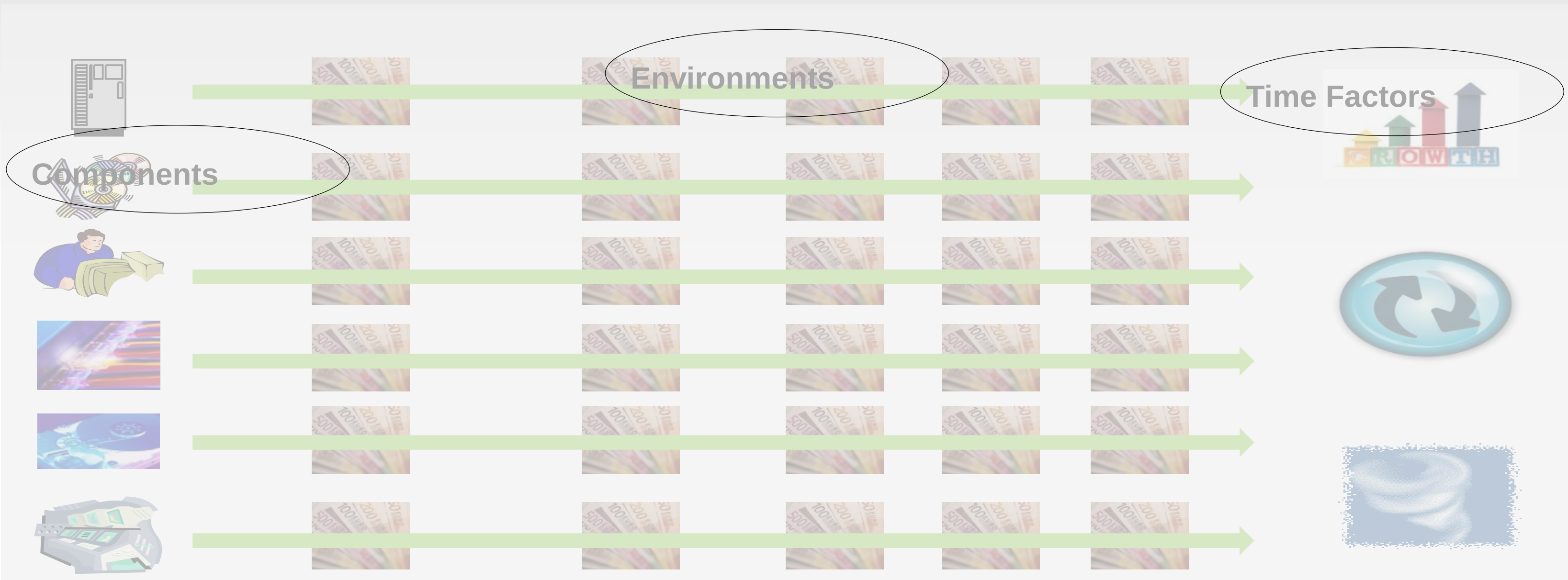
Environments Multiply Components



Time Factors Drive Growth And Cost

- Migration time and effort
- Business organic growth and/or planned business changes affect capacity requirements
  - e.g. Change of access channel or adding a new internet accessible feature can double or triple a components workload
  - Link a business metric (e.g. active customer accounts) to workload (e.g. daily transactions) and then use business inputs to drive the TCO case
- Other periodic changes – hardware refresh or software remediation

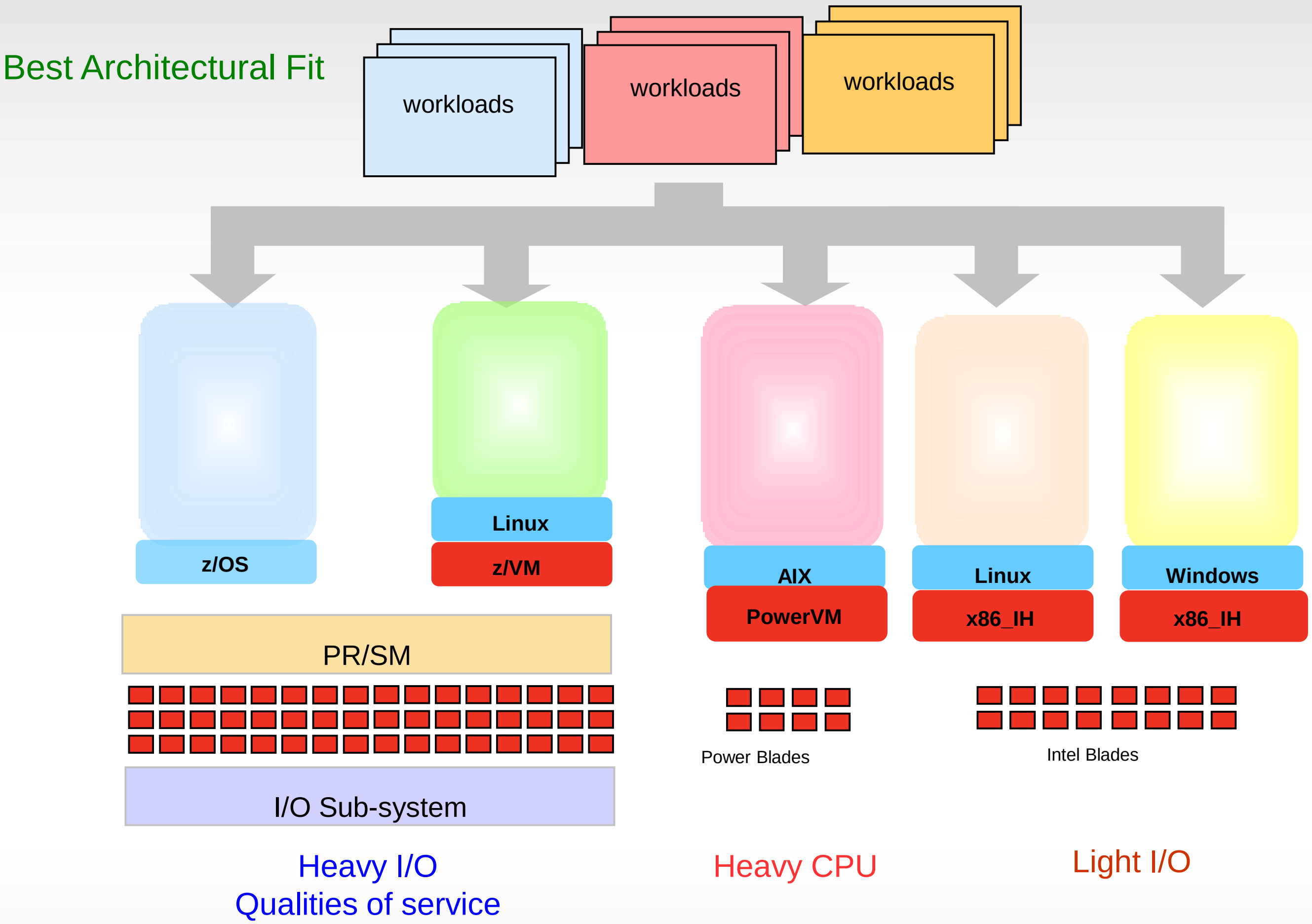
Non-Functional Requirements Can Drive Additional Resource Requirements



Availability … Security … Resiliency … Scalability …

Qualities of Service, Non-Functional Requirements

Workload Characteristics Influence The Best Fit Deployment Decision



Deploy or consolidate workloads on the environment best suited for each workload to yield lowest cost

Deploying Stand Alone Workloads With Heavy CPU Requirements

Benchmark to determine which platform provides the lowest TCA over 3 years

Heavy CPU workloads

- **IBM WebSphere ND**
- **Monitoring software**
- **On 8 core Nehalem servers**

Online banking workloads, each driving **460** transactions per second with light I/O

2 workloads per Intel blade

Scale to 16 cores

Virtualized on Intel 16 core HX5 Blade

$200,055 per workload

Best Fit

1 workload per POWER7 blade

PowerVM on PS701 8 core POWER7 Blade

$216,658 per workload

10 workloads per 32-way z/VM

z/VM on z196 CPC 32 IFLs

$328,477 per workload

Consolidation ratios derived from IBM internal studies. HX5 2.13GHz 2ch/16co performance projected from x3550 2.66GHz 2ch/12co measurements. zBX with x blades is a statement of direction only. Results may vary based on customer workload profiles/characteristics. Prices will vary by country.

Deploying Stand Alone Workloads With Light CPU Requirements

Benchmark to determine which platform provides the lowest TCA over 3 years

Light workloads

- **IBM WebSphere ND**
- **Monitoring software**
- **On 4 core "older" Intel**

Online banking workloads, each driving **22** transactions per second with moderate I/O

47 workloads per Intel blade

Virtualized on Intel
16 core HX5 Blade

$8,165 per workload

28 workload per POWER7 blade

Fast low cost threads

PowerVM on PS701
8 core POWER7 Blade

$7,738 per workload

Best Fit

155 workloads per 32-way z/VM

z/VM on z196 CPC
32 IFLs

$21,192 per workload

Consolidation ratios derived from IBM internal studies. HX5 2.13GHz 2ch/16co performance projected from x3550 2.66GHz 2ch/12co measurements. zBX with x blades is a statement of direction only. Results may vary based on customer workload profiles/characteristics. Prices will vary by country.

Deploying Stand Alone Workloads With Heavy I/O Requirements

IBM

Benchmark to determine which platform provides the lowest TCA over 3 years

Heavy I/O workloads

- **IBM WebSphere ND**
- **Monitoring software**
- **On 4 core "Older" Intel**

Online banking workloads, each driving **22** transactions per second, with **1 MB I/O per transaction**

1 workload
per Intel blade

Virtualized on Intel
16 core HX5 Blade

$400,109 per workload

1 workload
per POWER7 blade

PowerVM on PS701
8 core POWER7 Blade

$216,658 per workload

40 workloads
per 32-way z/VM

z/VM on z196 CPC
32 IFLs

$82,119 per workload

Best Fit

Consolidation ratios derived from IBM internal studies. HX5 2.13GHz 2ch/16co performance projected from x3550 2.66GHz 2ch/12co measurements. zBX with x blades is a statement of direction only. Results may vary based on customer workload profiles/characteristics. Prices will vary by country.

Oracle Coherence reduces TCA for read-only severe *sticky finger with think-time user* mobile workloads by 57% (forcing cache update)

Which platform provides the lowest TCA over 3 years?

Mobile Users

- **500 concurrent connections**
- **20 reads/session with 100ms think time (forcing a cache refresh)**
- **1 second cache invalidation (WXS scenario)**

Mobile read-only workload driving minimum throughput of **5,200** transactions per second and response time of 5ms

Liberty | CICS DB2

z/VM | **z/OS**

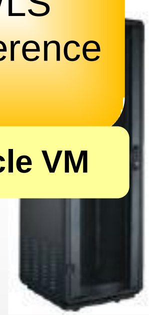
zEC12-705 + 3 IFLs

$21.8M (3 yr. TCA) Prod

$28.5M (3 yr. TCA) Prod+Dev/QA+DR

Oracle* WLS Coherence | CICS DB2

Oracle VM | **z/OS**



Exalogic X4-2 1/8th Rack (24 cores pro-rated)

zEC12-407

$8.6M (3 yr. TCA) Prod

$12.3M (3 yr. TCA) Prod+Dev/QA+DR

57% lower cost!

* Oracle Coherence performance projected from WXS Caching Test

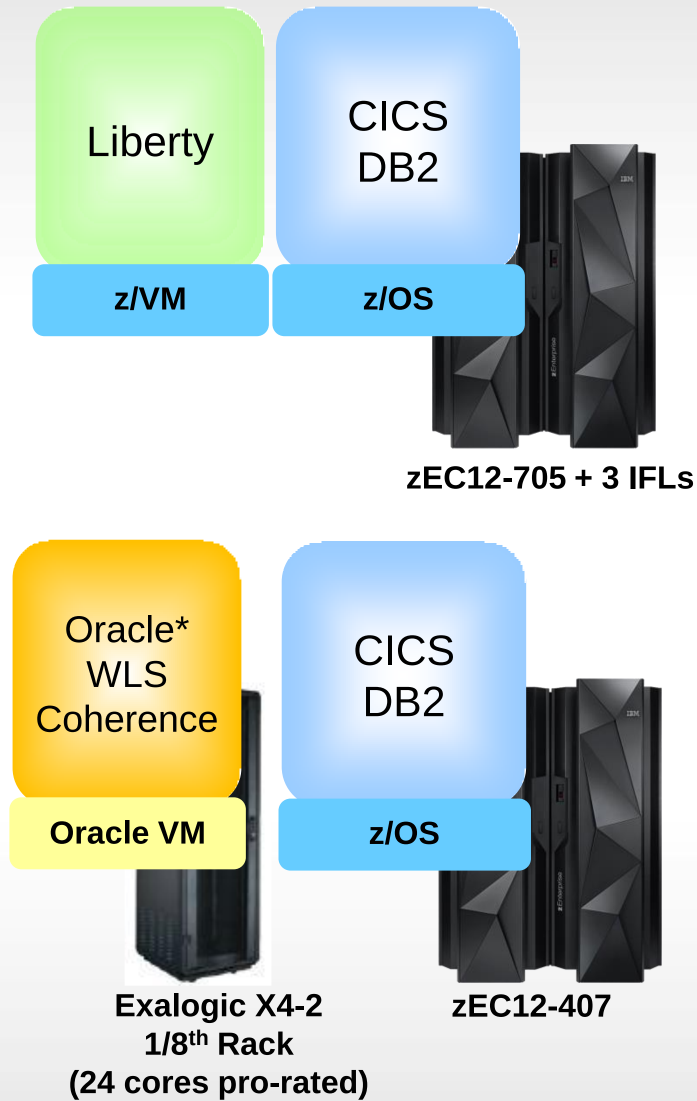
Oracle Coherence reduces TCA for read-only severe *sticky finger with think-time user* mobile workloads by 16% (forcing cache update) – using Mobile Workload Pricing

Which platform provides the lowest TCA over 3 years?

Mobile Users

- **500 concurrent connections**
- **20 reads/session with 100ms think time (forcing a cache refresh)**
- **1 second cache invalidation (WXS scenario)**

Mobile read-only workload driving minimum throughput of **5,200** transactions per second and response time of 5ms



$11.2M (3 yr. TCA) Prod

$14.7M (3 yr. TCA) Prod+Dev/QA+DR

$8.6M (3 yr. TCA) Prod

$12.3M (3 yr. TCA) Prod+Dev/QA+DR

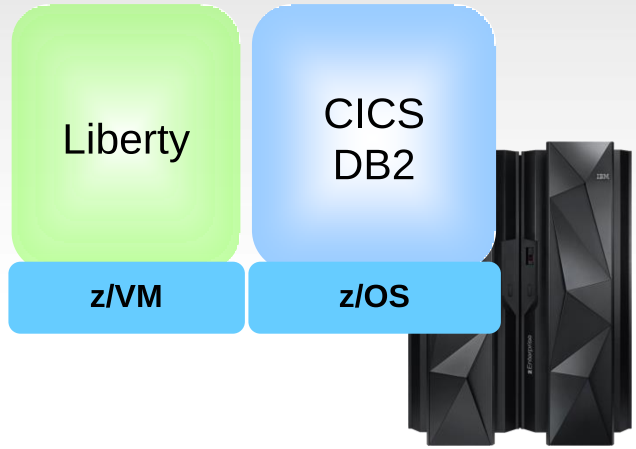
16% *lower cost!*

* Oracle Coherence performance projected from WXS Caching Test

Oracle Coherence reduces TCA for read-only moderate *sticky finger with think-time user* mobile workloads by 45% (forcing cache update)

Which platform provides the lowest TCA over 3 years?

Mobile Users



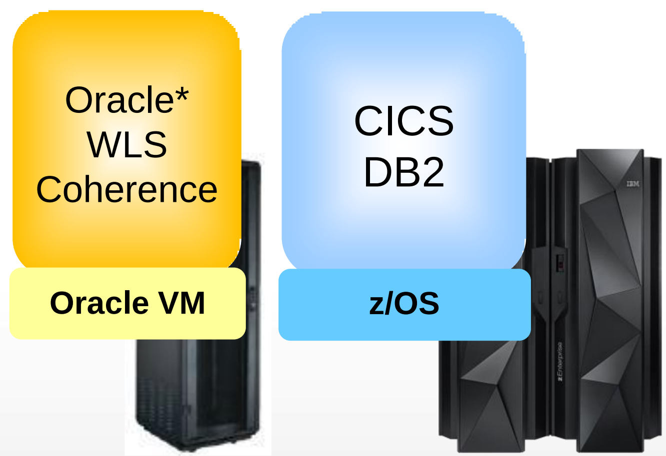
zEC12-703 + 3 IFLs

$16.3M (3 yr. TCA)
Prod

$21.3M (3 yr. TCA)
Prod+Dev/QA+DR

- **500 concurrent connections**
- **10 reads/session with 200ms think time (forcing a cache refresh)**
- **1 second cache invalidation (WXS scenario)**

Mobile read-only workload driving minimum throughput of **3400** transactions per second and response time of 2ms



Exalogic X4-2
1/8th Rack
(18 cores pro-rated)

zEC12-407

$8.4M (3 yr. TCA)
Prod

$11.8M (3 yr. TCA)
Prod+Dev/QA+DR

45%
lower cost!

* Oracle Coherence performance projected from WXS Caching Test

Oracle Coherence increases TCA by 5% for read-only moderate *sticky finger with think-time user* mobile workloads (forcing cache update) – using Mobile Workload Pricing

Which platform provides the lowest TCA over 3 years?

Mobile Users

- **500 concurrent connections**
- **10 reads/session with 200ms think time (forcing a cache refresh)**
- **1 second cache invalidation (WXS scenario)**

Mobile read-only workload driving minimum throughput of **3400** transactions per second and response time of 2ms

Liberty | CICS DB2

z/VM | **z/OS**

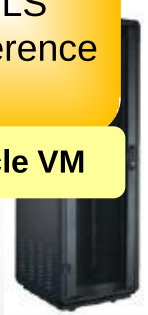
zEC12-703 + 3 IFLs

$8.5M (3 yr. TCA) Prod

$11.2M (3 yr. TCA) Prod+Dev/QA+DR

Oracle* WLS Coherence | CICS DB2

Oracle VM | **z/OS**



Exalogic X4-2 1/8th Rack (18 cores pro-rated)

zEC12-407

$8.4M (3 yr. TCA) Prod

$11.8M (3 yr. TCA) Prod+Dev/QA+DR

5% higher cost!

* Oracle Coherence performance projected from WXS Caching Test

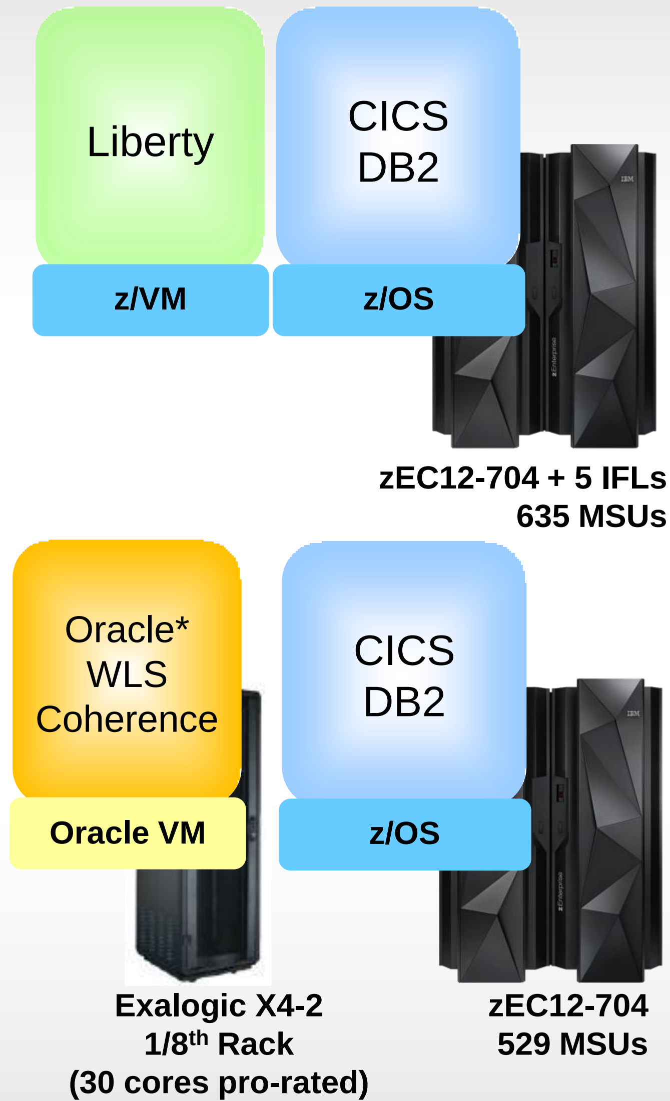
Using Oracle Coherence on Exalogic increases TCA by 5% for read-only *blended* workloads

Which platform provides the lowest TCA over 3 years?

Mobile Users

- **500 concurrent connections**
- **70% do 1 read/session; 25% do 4 reads/session; 5% do 20 reads/session with 100ms think time**
- **1 second cache invalidation (WXS scenario)**

Mobile read-only workload driving minimum throughput of **6,300** transaction per second and response time of 12ms



$19.8M (3 yr. TCA)
Prod

$25.9M (3 yr. TCA)
Prod+Dev/QA+DR

$19.9M (3 yr. TCA)
Prod

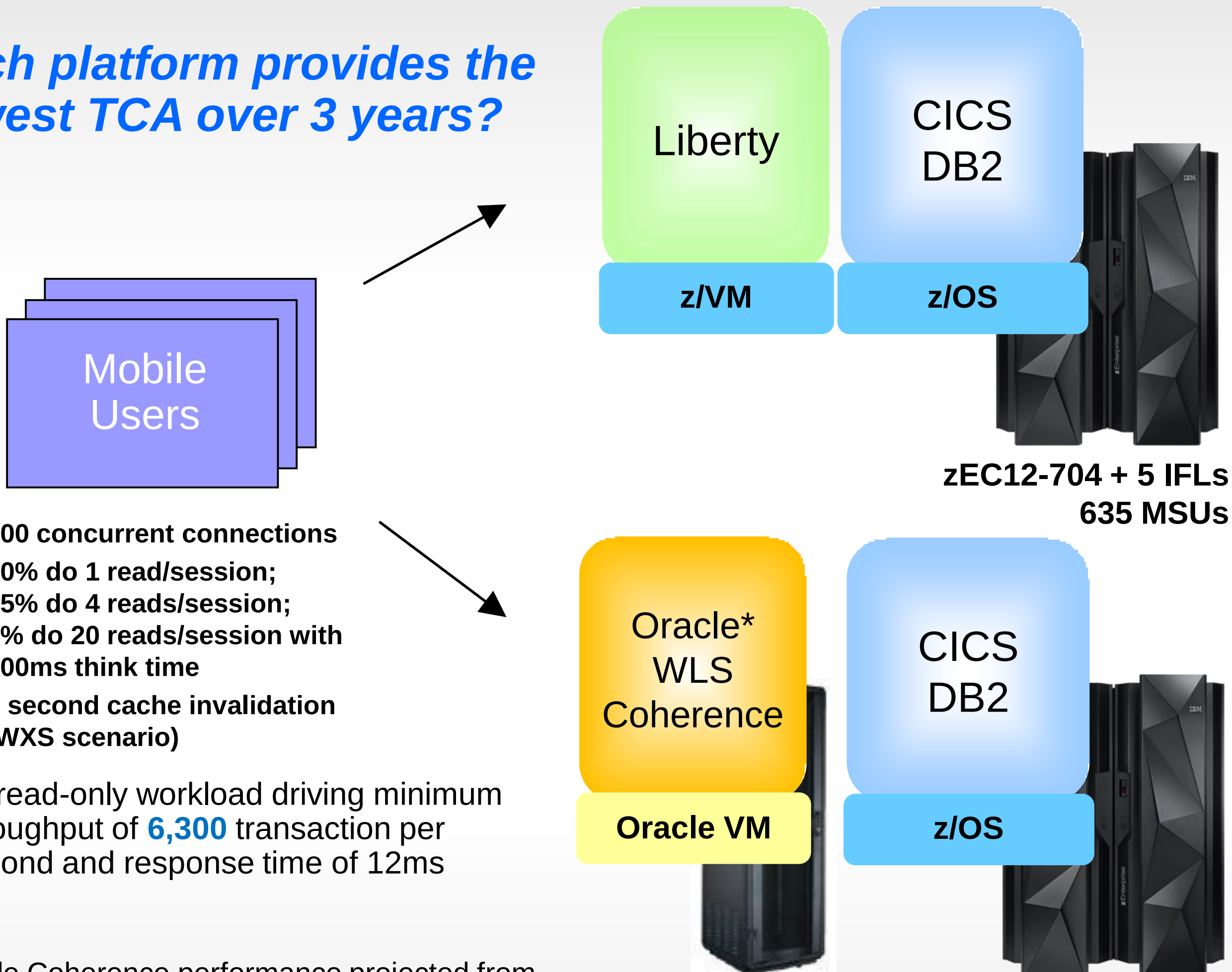
$27.2M (3 yr. TCA)
Prod+Dev/QA+DR

5% higher cost!

* Oracle Coherence performance projected from WXS Caching Test

Using Oracle Coherence on Exalogic increases TCA by 99% for read-only *blended* workloads – using Mobile Workload Pricing

Which platform provides the lowest TCA over 3 years?



- **500 concurrent connections**
- **70% do 1 read/session; 25% do 4 reads/session; 5% do 20 reads/session with 100ms think time**
- **1 second cache invalidation (WXS scenario)**

Mobile read-only workload driving minimum throughput of **6,300** transaction per second and response time of 12ms

* Oracle Coherence performance projected from WXS Caching Test

Exalogic X4-2
1/8th Rack
(30 cores pro-rated)

zEC12-704
529 MSUs

$10.4M (3 yr. TCA)
Prod

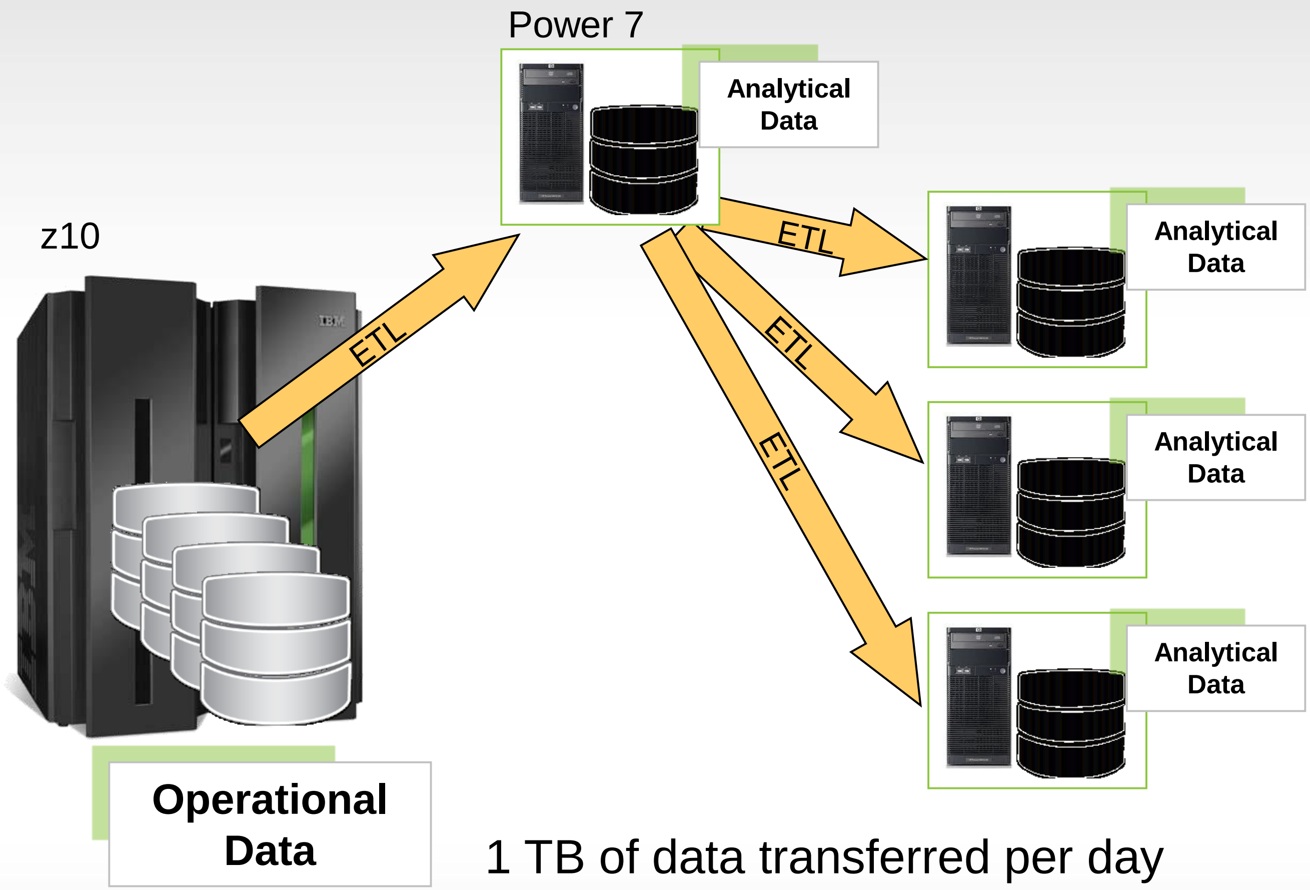
$13.7M (3 yr. TCA)
Prod+Dev/QA+DR

$19.9M (3 yr. TCA)
Prod

$27.2M (3 yr. TCA)
Prod+Dev/QA+DR

99% higher cost!

Observed ETL Cost Break Out TCA Plus TCO



1 TB of data transferred per day – one initial copy, plus three derivative copies

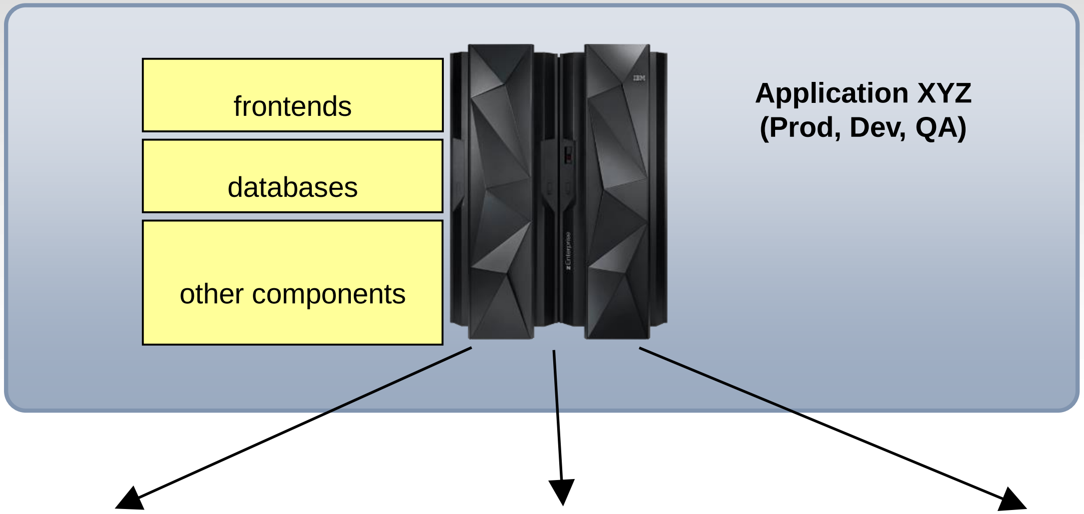
4 yr. amortized cost summary

System z Extract and Send	$2,861,600
Distributed Receive and Load	$4,466,140
Network	$430,408
System z Storage	$49,330
Distributed Storage	$238,720
System z Admin	$22,207
Distributed Admin	$143,090
System z Storage Admin	$5,880
Distributed Storage Admin	$51,960

Source: CPO internal study. Assume dist. send and load is same cost as receive and load.. Also, assume 2 switches and 2 T3 WAN connections.

What Happens In a TCO Study?

Workload identified for analysis



Deployment Choices

Do nothing

Optimize current environment

Deploy on other platforms

Key steps in analysis

1. Establish equivalent configurations
- Needed to deliver workload

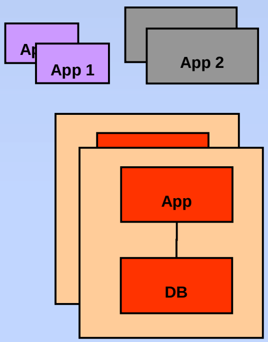
2. Compare Total Cost of Ownership
- TCO looks at different dimensions of cost

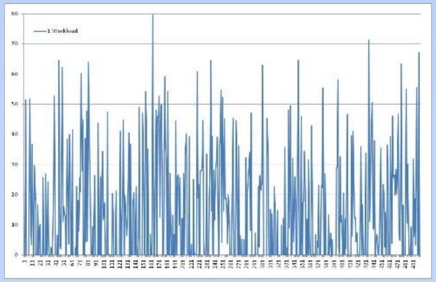
Approaches To Establishing Equivalent Configurations

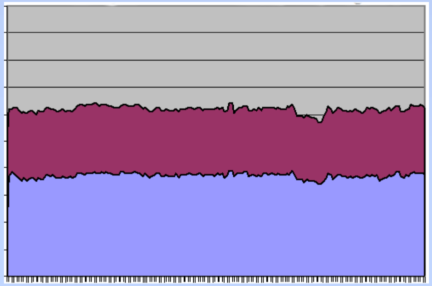
- Bottom up approach
  - Atomic benchmarks
  - Counting cycles, CPI comparisons …
  - IO, memory, cache, co-location effects …
  - Tends to show smaller core expansion factors
- Top down approach
  - “Real world” observations
  - Tends to show much larger core expansion factors
- When atomic benchmarks are assembled to represent “real world”, bottom up numbers approach top down numbers

How Can We Determine Equivalent Configurations?

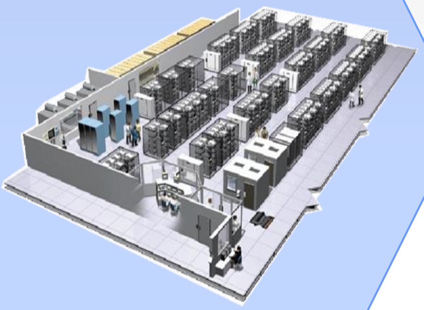
Real world aspects determine accurate equivalence







....



Platform factors

GHz, CPI, IO, co-location etc

Variability in demand

Different size servers

Workload Management

Mix workloads with different priorities

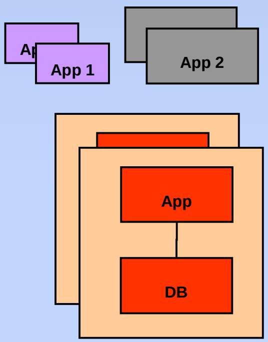
....

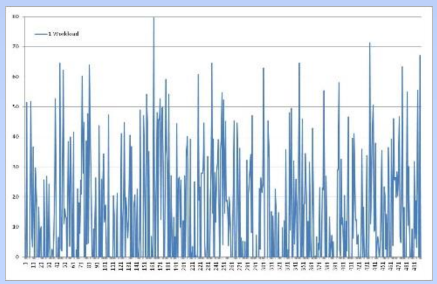
Top Down approach

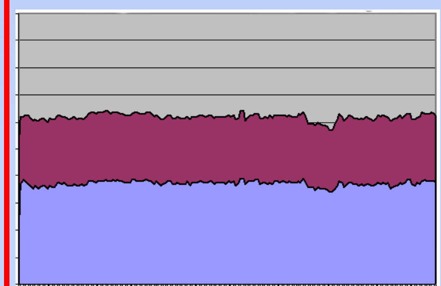
What we see in customer environments

How Can We Determine Equivalent Configurations?

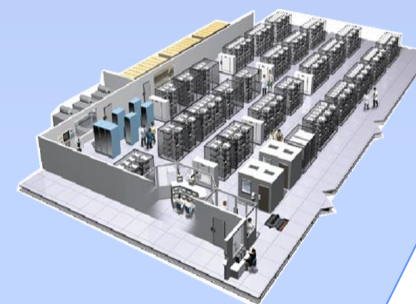
Real world aspects determine accurate equivalence







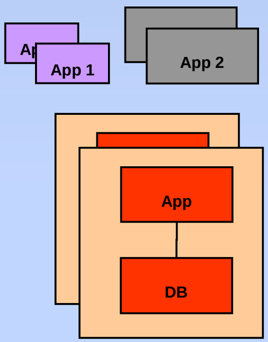
....

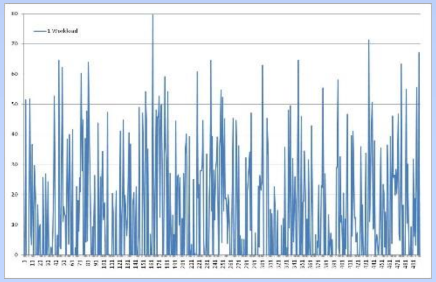


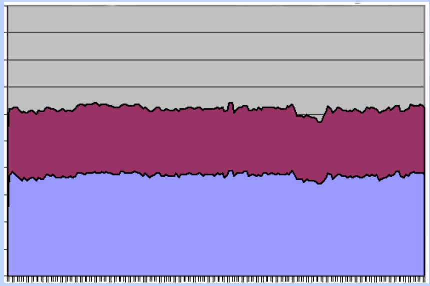
Platform factors GHz, CPI, IO, co-location etc	**Variability in demand** Different size servers	**Workload Management** Mix workloads with different priorities	....	**Top Down approach** What we see in customer environments

How Can We Determine Equivalent Configs?

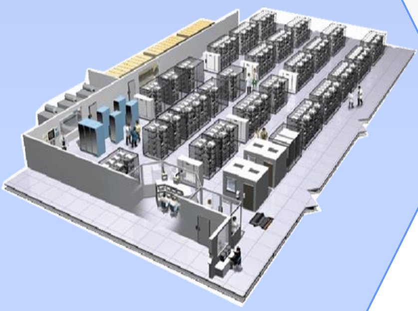
Real world aspects determine accurate equivalence







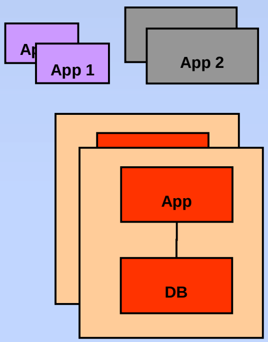
....

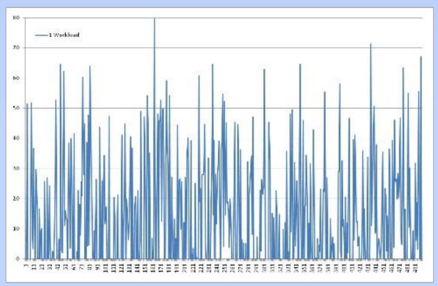


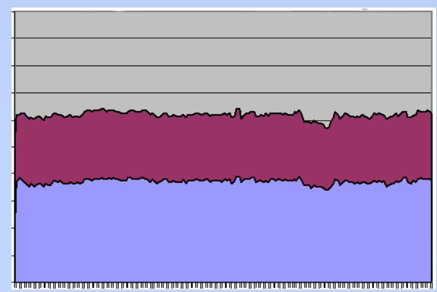
Size of the workload	**Variability in demand**	**Workload Management**	....	**Top Down approach**
Same software on Same size servers	Different size servers	Mix workloads with different priorities		What we see in customer environments

How Can We Determine Equivalent Configs?

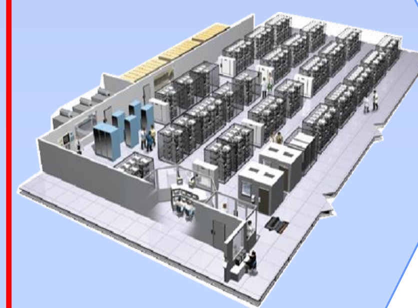
Real world aspects determine accurate equivalence







....



Size of the workload

Same software on
Same size servers

Variability in demand

Different size servers

Workload Management

Mix workloads
with different priorities

....

Top Down approach

What we see in customer environments

Core Proliferation For A Mid-sized Workload

6x 8-way HP DL Production / Dev
2x 64-way p595 Production / Dev
Application/MQ/DB2/Dev partitions

2x z900 3-way Production / Dev / QA / Test

6 processors *(1,660 MIPS)*

176 processors
(800,072 Performance units)

***29x** more cores!*

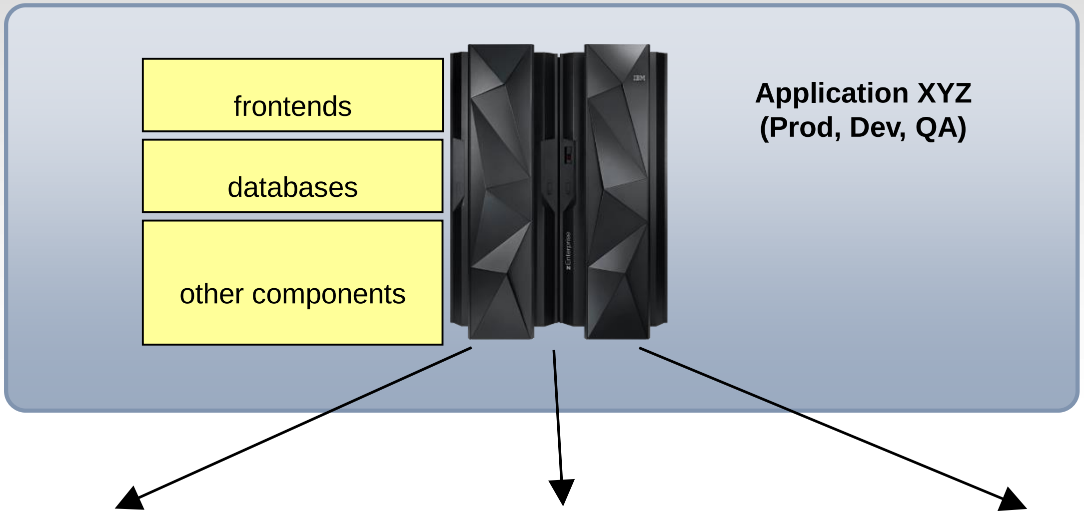
482 Performance Units per MIPS

So What Were The Total Costs In The Core Proliferation Cases We Saw Earlier?

Case	RPE/MIPS	Z Total Cost	Distributed Total Cost	Factor
Large benchmark	95	**$111M** (5 yr. TCA)	**$180M** (5 yr. TCA)	1.62x
Mid size offload	482	**$17.9M** (5 yr. TCO)	**$25.4M** (5 yr. TCO)	1.42x
Small offload	670	**$4.9M** (4 yr. TCO)	**$17.9M** (4 yr. TCO)	3.65x
Even smaller offload	499	**$4.7M** (5 yr. TCO)	**$8.1M** (5 yr. TCO)	1.72x

What Happens In a TCO Study?

Workload identified for analysis



Deployment Choices

Do nothing | **Optimize current environment** | **Deploy on other platforms**

Key steps in analysis

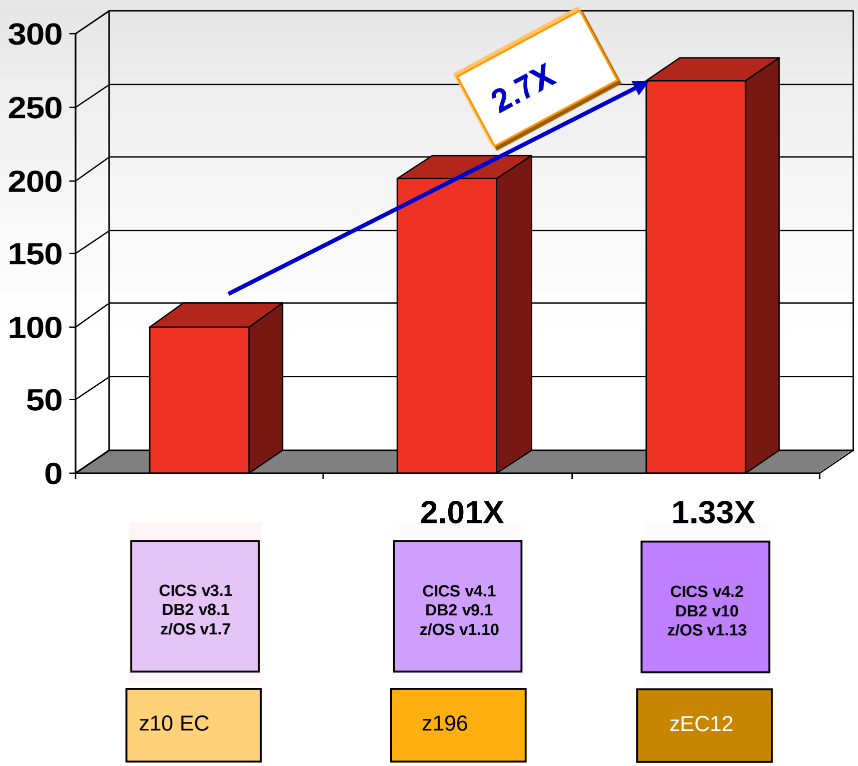
1. Establish equivalent configurations
- Needed to deliver workload

2. Compare Total Cost of Ownership
- TCO looks at different dimensions of cost

Lessons Learned Can Be Grouped Into Three Broad Categories

- Always compare to an optimum System z environment
- Look for not-so-obvious distributed platform costs to avoid
- Consider additional platform differences that affect cost

Performance Improvements Can Lower MLC Costs And Free Up Hardware Capacity



Customer examples:

(1) Large MEA bank

- Delayed upgrade from z/OS 1.6 because of cost concerns
- When finally did upgrade to z/OS 1.8
  - Reduced each LPAR's MIPS by 5%
  - Monthly software cost savings paid for the upgrade almost immediately

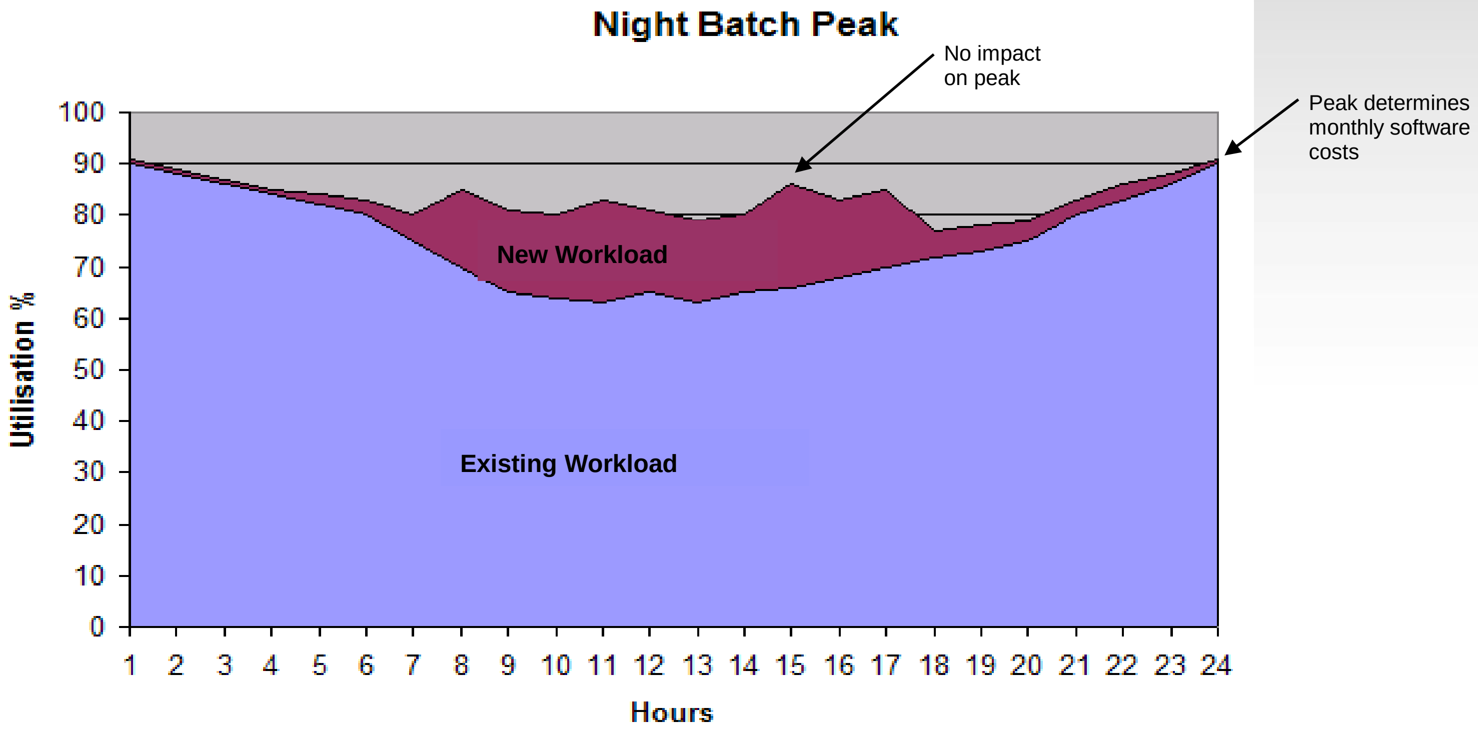
(2) Large European Auto company

- Upgraded to DB2 10
- Realized 38% pathlength reduction for their heavy insert workload
  - Other DB2 10 users saw 5-10% CPU reduction for traditional workloads

Additionally, save costs by moving to newer compilers and tuning

IBM internal core banking workload (Friendly Bank). Results may vary.

Sub-Capacity May Produce Free Workloads



- Standard “overnight batch peak” profile – drives monthly software costs
- Hardware and software are free for new workloads using the same middleware (e.g. DB2, CICS, IMS, WAS, etc.)
- Ensure you exploit any free workload opportunities, and conversely, avoid offloading free applications!

Leverage Accelerators Where Relevant

IBM

Standalone Pre-integrated Competitor V3

Quarter Unit

Unit Cost
$51/Reports per Hour

Workload Time	141 mins
Reports per Hour	68,581
Total Cost (3 yr. TCA) (HW+SW+Storage)	$3,530,041

IBM zEnterprise Analytics System 9700

IBM DB2 Analytics Accelerator (with PDA N2001-10)

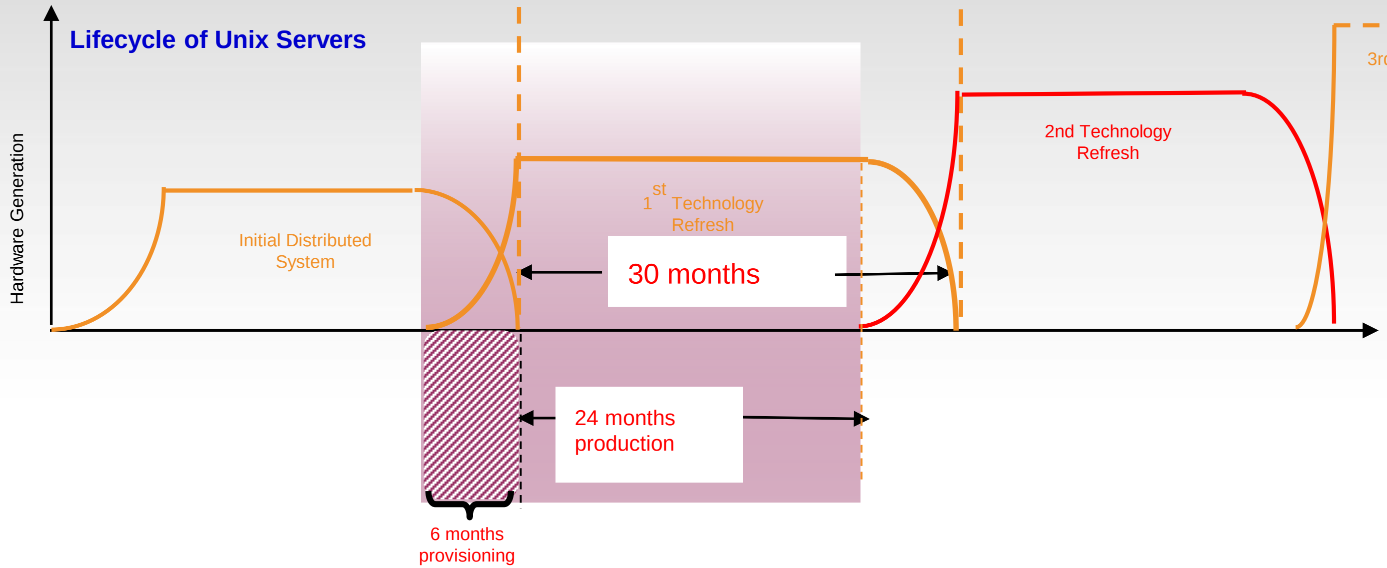
Unit Cost
$17/Reports per Hour

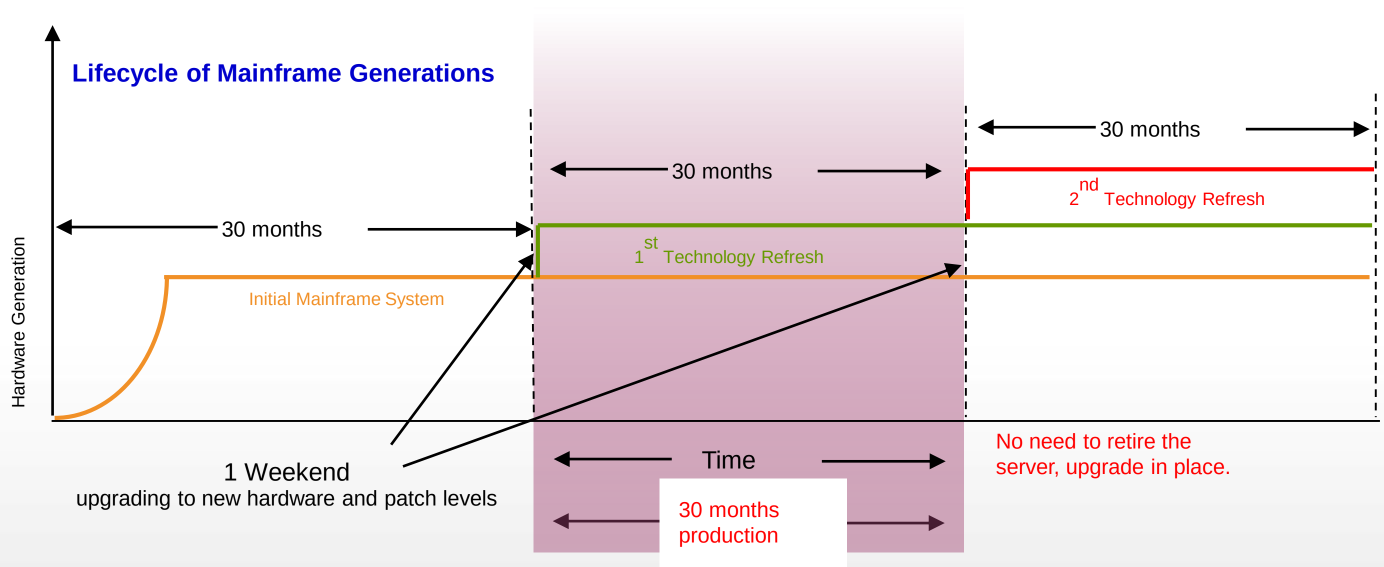
Workload Time	25 mins
Reports per Hour	386,798
Total Cost (3 yr. TCA) (13 GP + 12 zIIP, HW+SW+ Storage + Accelerator V3.1 with PDA N2001-10 hardware)	$6,464,849

***3x** price performance!*

Source: Customer Study on 1TB BIDAY data running 161,166 concurrent reports. Intermediate and complex reports automatically redirected to IBM DB2 Analytics Accelerator for z/OS. Results may vary based on customer workload profiles/characteristics. Note: Indicative 9700 pricing only internal to IBM, quotes to customer require a formal pricing request with configurations.

Distributed Servers Need To Be Replaced Every 3 To 5 Years





Refresh is normally even worse than just re-purchasing existing capacity as this real customer demonstrates:

Non-mainframe systems must co-exist for months at a time while being refreshed, requiring space, power, licenses etc. In this case only 24 months of productive work is realized for each 30 month lease period and the leases overlap up to 6 months

The mainframe by contrast is upgraded over a weekend and is fully productive at all times

Disaster Recovery On System z Costs Much Less Than On Distributed Servers

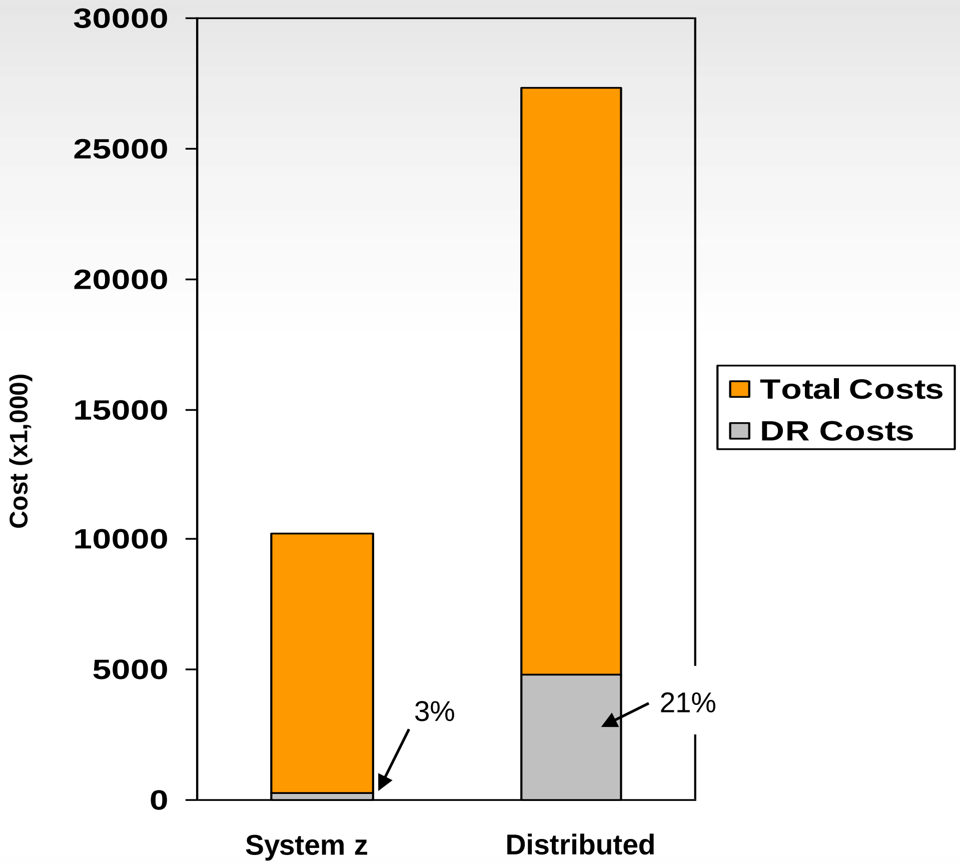
A large European insurance company with mixed distributed and System z environment at :

Disaster Recovery Cost as a percentage of Total Direct Costs:

System z – **3%**

Distributed – **21%**

Two mission-critical workloads on distributed servers had DR cost > 40% of total costs



Disaster Recovery Testing Is Typically More Expensive On Distributed Platforms Too

- A major US hotel chain
  - ~ 200 Distributed Servers (LinTel, Wintel, AIX, and HP-UX)

	Person-hours	*Elapsed days*	*Labor Cost*
Infrastructure Test (7 times)	1,144	7	$89,539
Full Test (4 times)	2,880	13	$225,416
Annual Total – Distributed	14,952*	73	$1,170,281
Mainframe Estimate	2,051*	10	$160,000

* Does not include DR planning and post-test debriefing

- Customer Recovery Time Objective (RTO) estimates:
  - Distributed ~ 48 hours to 60 hours
  - Mainframe ~ 2 hours

- Conclusion: Mainframe both simplifies and improves DR testing

Large Systems With Centralized Management Deliver Better Labor Productivity

Large US Insurance Company

HP Servers + ISV

Production Servers

HP 9000 Superdome RP4440

HP Integrity RX6600

Dev/Test Servers

HP 9000 Superdome RP5470

HP Integrity RX6600

Claims per year **327,652**

$0.12 per claim

$0.79 per claim

Mainframe support staff has 6.6x better productivity

IBM System z CICS/DB2

Total MIPS **11,302**

MIPS used for commercial claims processing prod/dev/test **2,418**

Claims per year **4,056,000**

Accumulated Field Data For Labor Costs

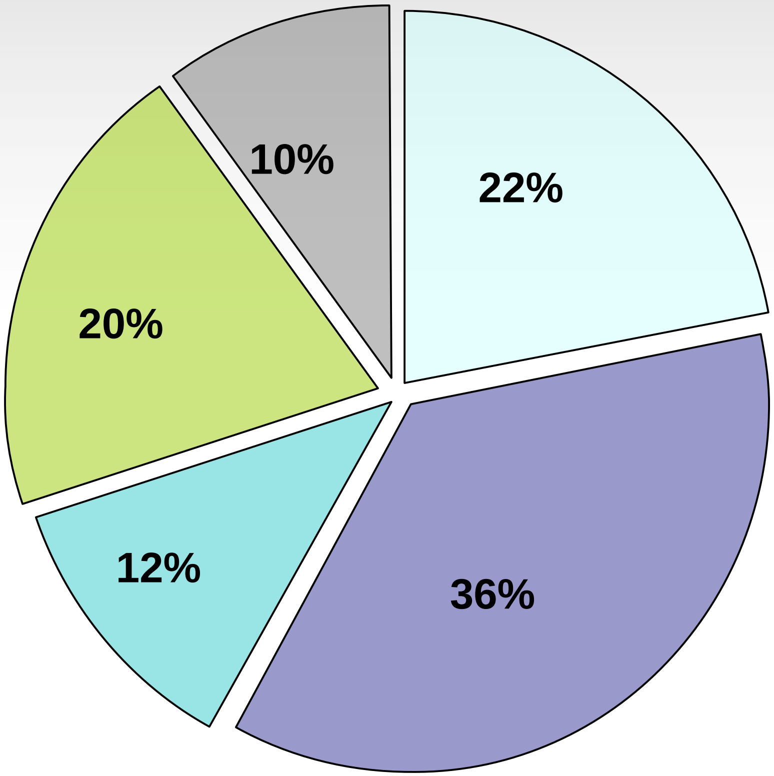
- Average of quoted infrastructure labor costs
  - **30.7** servers per FTE (dedicated Intel servers)
    - **67.8** hours per year per server for hardware and software tasks
  - **52.5** Virtual Machines per FTE (virtualized Intel servers)
    - **39.6** hours per year per Virtual Machine for software tasks and amortized hardware tasks
    - Typical 8 Virtual Machines per physical server

- Best fit data indicates
  - Hardware tasks are **32** hours per physical server per year
    - Assume this applies to Intel or Power servers
    - Internal IBM studies estimate 320 hours per IFL for zLinux scenarios
  - Software tasks are **36** hours per software image per year
    - Assume this applies to all distributed and zLinux software images

Labor model based on customer data from IBM studies

Five Key IT Processes For Infrastructure Administration

Time spent on each activity



- **Deployment Management**
  - *Hardware set-up and software deployment*
- **Incident/Capacity Management**
  - *Monitor and respond automatically*
- **Asset Management**
  - *Hardware and software asset tracking*
- **Security Management**
  - *Access control*
- **Change Management**
  - *Hardware and software changes*

Allocation based on customer data from IBM study

zManager Labor Cost Reduction Benefits Case Study

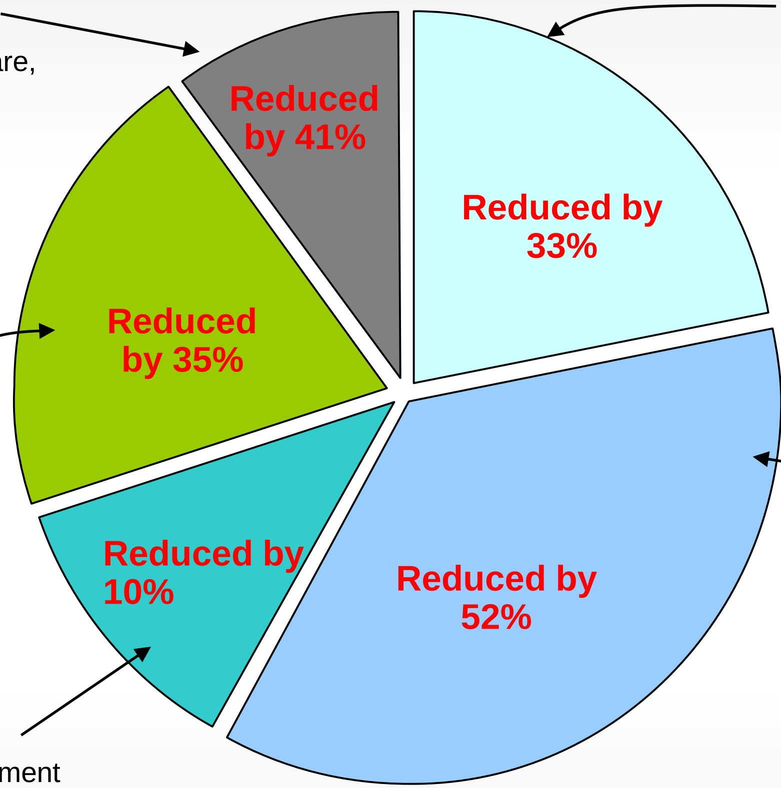
5032 total hours per year reduced by 38% to 3111 hours per year

Change Management

Standardization of images and firmware, visibility into relationships among resources

Deployment Management

Automatic setup and configuration of the hypervisor and out-of-the-box networks



Security Management

Centralized fine-grain administrator access control

Incident/Capacity Management

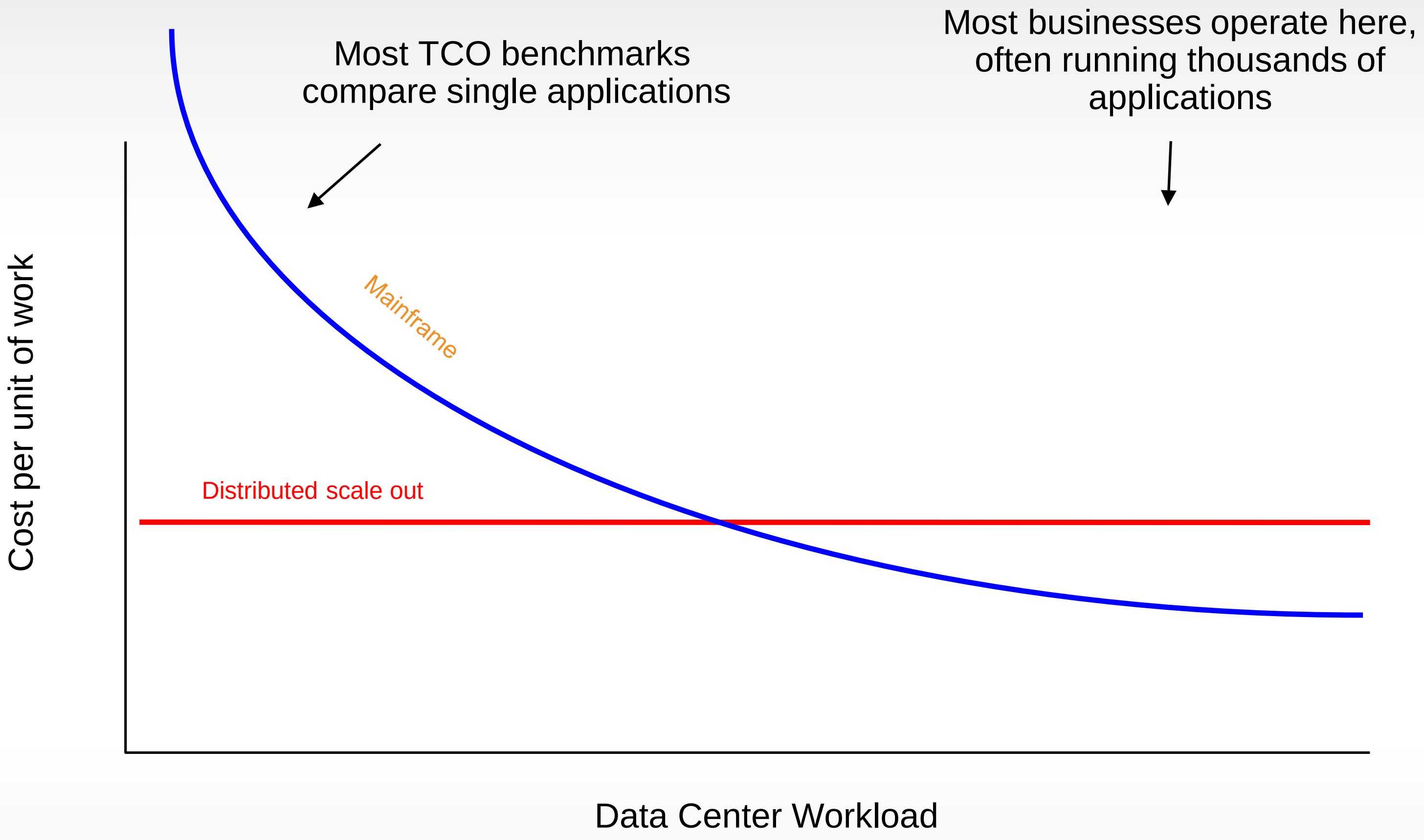
Automation to isolate and fix issues

Asset Management

Automated discovery, entitlement management

TCO: Understand The Complete Picture

Mainframe Cost/Unit of Work Decreases as Workload Increases



Cost Ratios in all TCO Studies

Average Cost Ratios (z vs Distributed)

		z	Distributed	z vs distributed (%)
Offload	**5-Year TCO**	**$16,351,122**	**$31,916,262**	**51.23%**
	Annual Operating Cost	$2,998,951	$4,405,510	68.07%
	Software	$10,932,610	$16,694,413	65.49%
	Hardware	$3,124,013	$3,732,322	83.70%
	System Support Labor	$3,257,810	$4,429,166	73.55%
	Electricity	$45,435	$206,930	21.96%
	Space	$59,199	$154,065	38.42%
	Migration	$438,082	$10,690,382	4.10%
	DR	$854,266	$2,683,652	31.83%
	Average MIPS	3,954		
	Total MIPS	217,452		
Consolidation	**5-Year TCO**	**$5,896,809**	**$10,371,020**	**56.86%**
	Annual Operating Cost	$716,184	$1,646,252	43.50%
	Software	$2,240,067	$6,689,261	33.49%
	Hardware	$2,150,371	$1,052,925	204.23%
	System Support Labor	$1,766,403	$2,395,693	73.73%
	Electricity	$129,249	$365,793	35.33%
	Space	$84,033	$205,860	40.82%
	Migration	$678,449	$0	
	DR	$354,735	$411,408	86.22%
	Average MIPS	10,821		
	Total MIPS	292,165		

Thank you.

Core Proliferation For A Very Large Workload

Configurations for equivalent throughput (10,716 Transactions Per Second)

16x 32-way HP Superdome
App. Production / Dev / Test

8x 48-way HP Superdome
DB Production / Dev /Test

zEC12 41-way Production / Dev / Test

***41* GP processors** *(38,270 MIPS)*

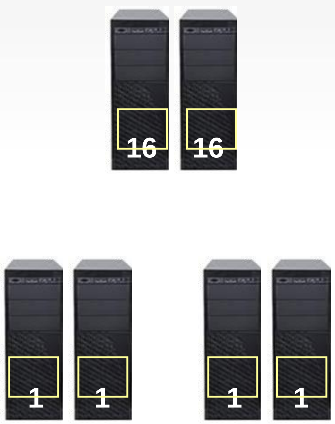
***896* processors** *(3,668,600 Perf Units)*

***22x** more cores!*

Core Proliferation For
A Small Offload Project

2x 16-way Production / Dev / Test / Education
App, DB, Security, Print and Monitoring
4x 1-way Admin / Provisioning / Batch Scheduling

z890 2-way Production / Dev / Test / Education
App, DB, Security, Print, Admin & Monitoring



***0.88* processors** *(332 MIPS)*

***36* Unix processors** *(222,292 Performance Units)*

41x more cores

Almost 5 Year Migration

670 Performance Units per MIPS

1 CICS region in production!!
CICS/IDMS migrated to CICS/DB2.
Accessing DB2 thru mapping layer

No Disaster Recovery

Core Proliferation For A Smaller Offload Project

z890 Production / Test

4x p550 (1ch/2co)
Application and DB

***0.24* processors** *(88 MIPS)*

***8* Unix processors**
(43,884 Performance Units)

33x more cores

3 Year Migration

499 Performance Units per MIPS

Just Completed x86 Offload

3x HP DL580 (2ch/20co)
Production / Dev / Test
(2011 x86 technology)

z800 Production /
Dev / Test
(2002 mainframe
technology)

2.1 ***processors*** *(499 MIPS)*

60 Linux processors
(383,022 Perf Units)

29x more cores

(despite the 9 year technology gap!)

1.5 Year Migration

768 Performance Units per MIPS